Instytut Badań Systemowych, Polska Akademia Nauk

Autoreferat pracy doktorskiej

Estymacja gęstości rozkładu niestacjonarnych danych strumieniowych

mgr inż. **Tomasz Rybotycki**

Instytut Badań Systemowych, Polska Akademia Nauk
Studia doktoranckie „Techniki informacyjne – teoria i zastosowania”

Promotor: prof. dr hab. inż. **Piotr Kulczycki**

Warszawa, 2023

1. WSTĘP

Postęp technologiczny w zakresie technik numerycznych, zarówno w dziedzinie sprzętu jak i oprogramowania, który miał miejsce w minionym czasie, umożliwia uzyskiwanie danych na szeroką skalę. Pojawia się przy tym szereg nowych problemów, które każdy projektant musi rozwiązać, przykładowo przechowywanie, przetwarzanie, analiza czy wydobywanie zawartej w nich wiedzy. W ciągu ostatniej dekady, znacząco wzrosło zainteresowanie szczególnym rodzajem danych, cechujących się nieograniczonością i ciągłością napływu – tak zwanymi danymi strumieniowymi. W praktyce dodatkowo są one niestacjonarne (ich rozkład ewoluuje w czasie), a zatem ich charakterystyki są zmienne, co jeszcze bardziej utrudnia analizę. Dane tego typu mogą ulegać dezaktualizacji. Nierzadko ważne są też wymagania z zakresu szybkości obliczeń. Co więcej, natura permanentnie napływających danych powoduje, że często zawierają one elementy nietypowe będące wynikiem błędów, aczkolwiek czasem reprezentujących nowo formujące się tendencje. Zwłaszcza ten drugi aspekt jest szczególnie warty konstruktywnego podkreślenia, gdyż wczesna wiedza o powstającym trendzie pozwala na odpowiednie wyprzedzenie działań koniecznych w zmiennej rzeczywistości. Ostatecznie, odpowiednia, spełniająca wymogi współczesnych zastosowań, analiza danych strumieniowych wymaga uwzględnienia szeregu istotnych czynników, często nieobecnych w klasycznych problemach przy ustalonym zbiorze danych. Czyni to tego typu analizę niezwykle cenną z aplikacyjnego punktu widzenia, ale też wymagającą z perspektywy badawczej.

Hipotezą poznawczą przedłożonej dysertacji, jak i jej celem badawczym, było rozstrzygnięcie możliwości opracowania algorytmu, przeznaczonego do estymacji gęstości rozkładu strumieni danych. Opracowany algorytm jest szczególnie predestynowany do przypadków niestacjonarnych, także o niestałej intensywności zmian, które mogą występować naprzemiennie z warunkami stacjonarnymi.

Struktura procedury ma charakter modułowy, przy czym każdy z tych modułów może być pominięty lub indywidualnie dostosowany do istniejących uwarunkowań. Charakterystyka badanego strumienia jest analizowana w ramach algorytmu, którego aktualne wartości parametrów ustalane są na bieżąco w trakcie działania. Koncepcja procedury oparta została na metodyce estymatorów jądrowych, co pozwala wyznaczyć gęstość dla dowolnego występującego w praktyce rozkładu danych. W przypadku wykrycia niestacjonarności wykorzystywane są elementy prognozowania statystycznego, co umożliwia niejako wyprzedzenie zachodzących zmian. Wykrywane są także elementy nietypowe, co w połączeniu z powyższym, pozwala określić, które z nich związane są z tendencjami wzrastającymi, a które z zanikającymi. Efektywność algorytmu została przebadana i pozytywnie zweryfikowana za pomocą ilustracyjnych danych syntetycznych oraz rzeczywistych pomiarów meteorologicznych.

2. NIESTACJONARNE DANE STRUMIENIOWE

Strumieniem danych $\{X_t\}_{t=1,2,\dots}$ nazywamy uporządkowaną sekwencję elementów

$$\{X_t\}_{t=1,2,\dots} = (x_1, x_2, \dots, x_t, \dots) \ , \tag{1}$$

gdzie $t \in \mathbb{N}\backslash\{0\}$ oznacza bieżącą wartość zmiennej niezależnej, najczęściej utożsamianej z czasem. Najczęściej zakłada się, że poszczególne elementy $x_t \in \mathbb{R}^n$ pojawiają się równomiernie dla kolejnych wartości zmiennej niezależnej t. Strumienie danych mają ponadto trzy zasadnicze cechy. Pierwszą z nich jest ogromna (potencjalnie nieskończona) liczność elementów. Kolejną – znaczna szybkość przybywania kolejnych danych. Strumienie najczęściej są też niestacjonarne, czyli ich rozkład w istotny sposób zmienia się w czasie (ang. *concept drift*).

Jedną z możliwości wnioskowania o stacjonarności strumienia danych dostarcza statystyczny test KPSS [9]. Jest on testem istotności; weryfikuje hipotezę stacjonarności procesu. Niech zatem dany będzie szereg czasowy $\{X_t\}_{t=1,2,\dots}$, który zostały wygenerowany z rzeczywistego ($n = 1$) procesu stochastycznego, podlegającego badaniom. Test KPSS sformułowany został w dwóch wersjach: zakładającą istnienie trendu lub nie. Poniżej użyta została druga opcja, a potencjalne trendy traktowane będą jako przejaw niestacjonarności strumienia.

Oznaczmy przez $KPSS$ wartość statystyki testowej testu KPSS. Wzorując się na aksjomatach logiki rozmytej, zdefiniowany został stopień niestacjonarności $sgmKPSS$ jako

$$sgmKPSS(KPSS) = sgm(0.995\ KPSS - 2.932) \ , \tag{2}$$

gdzie sgm oznacza funkcję sigmoidalną. Wartość powyższej funkcji $sgmKPSS$ należy zatem do przedziału [0, 1] i interpretowana będzie jako stopień niestacjonarności strumienia lub – uwypuklając interpretację z zakresu logiki rozmytej – stopień przynależności badanego strumienia do klasy strumieni niestacjonarnych. Współczynniki występującej we wzorze (2) funkcji liniowej zostały dobrane tak aby wartość krytyczna 0,739 odpowiadająca najmniejszemu – stosowanemu w praktyce – poziomowi istotności 0,01 została przyporządkowana największemu – stosowanemu w praktyce – poziomowi istotności 0,1, czyli

$$sgmKPSS(0{,}739) = 0{,}1 \ , \tag{3}$$

natomiast analogicznie najmniejsza wartość krytyczna 0,347 w 7-krotność najmniejszego poziomu krytycznego, a więc

$$sgmKPSS(0{,}347) = 7 \cdot 0{,}01 \ . \tag{4}$$

Widać zatem, że maksymalna i minimalna wartość krytyczna testu KPSS zmieniane są odpowiednio w maksymalny poziom krytyczny testu KPSS oraz 7-krotność minimalnego. Zamiana wartości minimalnej i maksymalnej wynika z faktu, że najmniejszemu poziomowi istotności w teście $KPSS$ odpowiada największa wartość krytyczna i odwrotnie. Wybór wzmiankowanej powyżej krotności 7 został dokonany na drodze heurystycznej, zainspirowany metodą Zieglera-Nicholsa, podstawową w klasycznej teorii regulacji z zakresu automatyki, używanej tam do wyznaczania nastaw sterowników PID. Dla krotności 7 występowało bowiem minimum w sensie wskaźnika całkowego błędu estymacji gęstości danych strumieniowych o rozkładzie normalnym, przy jednostokowym odchyleniu standardowym i wartości oczekiwanej w postaci skoku jednostkowego, czyli $N(0,1)\rightarrow N(1,1)$.

Stopień niestacjonarności został określony powyżej dla jednowymiarowych szeregów czasowych. W przypadku wielowymiarowym zdefiniujmy

$$sgmKPSS = \max_{j=1,2,\dots,n} sgmKPSS_j \ , \tag{5}$$

gdzie $sgmKPSS_j$ reprezentuje stopień niestacjonarności (2) wyznaczony dla j-tej współrzędnej. Takie ujęcie implikuje, że o niestacjonarności wielowymiarowego procesu stochastycznego stanowi niestacjonarność najszybciej zmieniającej się współrzędnej.

3. ESTYMACJA GĘSTOŚCI ROZKŁADU

Gęstością rozkładu n-wymiarowej zmiennej losowej nazywa się mierzalną funkcję $f\colon \mathbb{R}^n \rightarrow [0,\infty)$ taką, że dla każdego zbioru borelowskiego $B \subset \mathbb{R}^n$ zachodzi

$$P(B) = \int_B f(x)\,dx \ . \tag{6}$$

Załóżmy, że badany rozkład posiada gęstość. Jej jądrowy estymator $\hat{f}\colon \mathbb{R}^n \rightarrow [0,\infty)$ definiuje się wzorem

$$\hat{f}(x) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{K}(x, x_i, h) \ , \tag{7}$$

gdzie $m \in \mathbb{N}\backslash\{0,1,2\}$ oznacza liczbę empirycznie uzyskanych n-wymiarowych obserwacji $x_1, x_2, \dots, x_m \in \mathbb{R}^n$ na podstawie których budowany jest estymator, wektor $h \in \mathbb{R}^n$ o elementach dodatnich określa się mianem parametru wygładzania, natomiast funkcja $\mathcal{K}\colon \mathbb{R}^n \rightarrow [0,\infty)$ jest (n-wymiarowym) jądrem, wobec której formułuje się następujące założenia:

a. jest funkcją mierzalną;
b. jest symetryczna względem zera, czyli $\mathcal{K}(x) = \mathcal{K}(-x)$ dla $x \in \mathbb{R}^n$;

c. ma słabe maksimum globalne w zerze, a zatem $\mathcal{K}(0) \geq \mathcal{K}(x)$ dla $x \in \mathbb{R}^n$;

d. spełnia warunek jednostkowej całki $\int_{\mathbb{R}^n} \mathcal{K}(x)\mathrm{d}x = 1$.

Przy wielowymiarowej analizie proponowane ujęcie produktowe, gdzie jądro $\mathcal{K}$ definiuje się jako

$$\mathcal{K}(x) = \prod_{j=1}^{n} K_j(x_j) \ , \tag{8}$$

gdzie K_j oznacza jądro jednowymiarowe odpowiadające j-tej współrzędnej. W praktyce, jądra jednowymiarowe są jednakowe dla wszystkich współrzędnych; wówczas można zapisać $K_j \equiv K$. Wprowadzając parametr j wobec poszczególnych współrzędnych x, h oraz x_i, zdefiniowane we wzorze (8) jądro można zapisać w ostatecznej postaci:

$$\mathcal{K}(x, x_i, h) = \prod_{j=1}^{n} \frac{1}{h_j} K\left(\frac{x_j - x_{i,j}}{h_j}\right) \ . \tag{9}$$

Jednowymiarowe jądro K najczęściej dobiera się tak, aby estymator konstruowany przy jego pomocy miał cechy dogodne w kontekście badanego zagadnienia. Nie ma ono bowiem istotnego znaczenia w kontekście dokładności estymacji, zwłaszcza przy powszechnych obecnie dużych licznościach zbiorów. Zwykle stosowanym jest jądro normalne dane wzorem

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \ . \tag{10}$$

Podczas wnioskowania o niestacjonarności badanych strumieni, istotnym aspektem będzie uogólnienie wzoru (7) tak, aby możliwe było sukcesywne powiększanie lub zmniejszanie znaczenia poszczególnych obserwacji, przykładowo tych które w miarę upływu czasu stają się coraz mniej aktualne. Zdefiniować można zatem ważoną postać estymatora jądrowego

$$\hat{f}(x) = \frac{1}{\sum_{i=1}^{m} w_i} \sum_{i=1}^{m} w_i \, \mathcal{K}(x, x_i, h) \ , \tag{11}$$

gdzie $w_i \geq 0$ oznacza wagę (znaczenie) i-tej obserwacji, przy czym nie wszystkie wagi są równe zero. Łatwo można zauważyć, że jeśli wagi w_i są równe, to wzór (11) sprowadza się do zależności (7), a zatem formuła (11) jest istotnym uogólnieniem definicji (7).

W przeciwieństwie do wyboru typu jądra, wyznaczenie wartości parametru wygładzania ma istotne znaczenie dla dokładności i własności estymatora jądrowego. W niniejszej pracy wykorzystywana była metoda podstawień (ang. *plug-in*) [2]. Ma ona kwadratową złożoność obliczeniową względem liczności zbioru m, co przy stosowanych w algorytmie licznościach z zakresu od 100 do 1000 (por. sekcja 4.1) jest wystarczające z aplikacyjnego punktu widzenia.

4. OPIS ALGORYTMU

Procedura opracowana w ramach przedłożonej pracy została nazwana DEDSTA jako akronim angielskiej nazwy *Density Estimation for Data Stream with possible Trend*, przy czym ostatnia litera pochodzi od wyrazu *Algorithm*. Służy ona do estymacji gęstości rozkładu danych strumieniowych, przy czym została ukierunkowana na strumienie niestacjonarne, zwłaszcza w obecności trendu. Zakres stosowania obejmuje również przypadki, gdy niestacjonarność ma zmienny charakter i intensywność, a także może występować naprzemiennie z okresami stacjonarnymi. Dodatkowo opracowana procedura umożliwia identyfikację elementów nietypowych, z zaznaczeniem które z nich reprezentują tendencje nowopowstałe, a które wygasające.

Konstrukcja estymatora przy pomocy procedury DEDSTA jest modułowa i można ją scharakteryzować jako estymację gęstości rozkładu strumienia danych, w modelu ruchomego okna ze zmienną licznością i modyfikacją wag jego elementów. Określenie liczności okna zostanie przedstawione w sekcji 4.1, natomiast modyfikacja wag wprowadzana będzie w kolejnych sekcjach 4.2-4.4. W pierwszej z nich przedłożona jest dezaktualizacja „starszych" obserwacji. W przypadku wystąpienia trendu zostaje on zidentyfikowany metodami prognozowania statystycznego i wprowadzony do procedury DEDSTA w ramach sekcji 4.3. Pozwala to także określić które z elementów nietypowych są związane ze zjawiskami wstępującymi, a które z regresywnymi – zagadnieniu temu poświęcona jest sekcja 4.4.

4.1. Liczność okna

Procedura DEDSTA operuje na trzech różnych licznościach rezerwuaru, czyli ruchomego okna złożonego z ostatnich elementów badanego strumienia danych. Niech zatem dane będą trzy liczby $m_{min}, m, m_0 \in \mathbb{N}\backslash\{0,1\}$, takie że $m_{min} \leq m \leq m_0$. Określają one odpowiednio minimalną, aktualną i standardową (maksymalną) liczność rezerwuaru. Parametry m_{min} oraz m_0 są stałe, natomiast wartości m podlegają ciągłym zmianom, dopasowując się do aktualnych cech badanego strumienia danych.

Aktualną liczność rezerwuaru określa się w procedurze DEDSTA następującą zależnością:

$$m = \begin{cases} m_{min} & \text{gdy} \quad m_* < m_{min} \\ m_* & \text{gdy} \quad m_{min} \leq m_* \leq m_0 \\ m_0 & \text{gdy} \quad m_* > m_0 \end{cases}, \tag{12}$$

gdzie m_* wyznaczona jest jako

$$m_* = int[1.1\, m_0\, (1 - sgmKPSS)] \ , \tag{13}$$

przy czym $sgmKPSS$ reprezentuje stopień niestacjonarności badanego strumienia w danej chwili. Współczynnik 1.1 w powyższej zależności pozwala wyeliminować „ogon" statystyki $KPSS$, czyli obserwacji mniejszych od wartości krytycznej dla największego w praktyce poziomu istotności 0,1.

Po ustaleniu zależności (13) dodatkowo uwiarygodniono założoną heurystycznie wartość $m_0 = 1{,}000$. Ponownie odwołując się do zagadnienia minimalizacji błędu z użyciem normy całkowej w odpowiedzi na skok jednostkowy – znanego z klasycznej techniki regulacji – ustalono, że minimum to przeważnie występuje dla $m_0 \cong 1{,}000$. W przypadku rozkładów istotnie wielomodalnych – w oparciu o te same przesłanki – można polecić ewentualne nieznaczne zwiększenie tej wartości o 100 na każdy dodatkowy mod. Wartość parametru m_{min} powinna zależeć od największej prędkości zmian. W szczególności proponuje się

$$m_{min} = int\left(\frac{m_0}{10}\right) , \tag{14}$$

co przy proponowanych wyżej wartościach m_0 pozwala efektywnie śledzić zmiany nie szybsze niż $0{,}01 \cdot \hat{\sigma}$ na krok.

4.2. Dezaktualizacja obserwacji

Elementy rezerwuaru mogą podlegać sukcesywnej dezaktualizacji w miarę upływu czasu. Oznacza to, że informacja właściwa starszym elementom potencjalnie staje się coraz mniej wartościowa. Koncepcja ta realizowana będzie w procedurze DEDSTA poprzez określenie odpowiedniego wpływu formuły dezaktualizacji na wartości wag w_i poszczególnych elementów, na podstawie których wyznaczany jest estymator jądrowy (11); por. sekcja 4.5.

Zdefiniujmy następujące współczynniki charakteryzujące funkcję dezaktualizacji

$$w_i^* = 2\left[1 - \frac{\alpha(i-1)}{m}\right] \quad \text{dla } i = 1, 2, \ldots, m , \tag{15}$$

gdzie parametr $\alpha \in [0, 1]$ stanowić będzie o intensywności tego procesu. W przypadku stacjonarności warto zatem przyjąć wartość $\alpha = 0$, aby wszystkie obserwacje były tak samo ważne, sukcesywnie ją zwiększając w miarę wzrostu niestacjonarności do największej dopuszczalnej wartości 1. W naturalnej konsekwencji zostało przyjęte:

$$\alpha = sgmKPSS , \tag{16}$$

gdzie wartość $sgmKPSS$ dana jest wzorami (2) i (5) i charakteryzuje stopień niestacjonarności strumienia w danym momencie.

Liniowa postać formuły (15) została uwiarygodniona na podstawie przeprowadzonych symulacji

komputerowych. Nieliniowe postacie tej formuły – przykładowo kwadratowa czy też pierwiastkowa – nie dawały korzystniejszych rezultatów, jakby nadmiernie pomniejszając informację reprezentowaną przez dawne (postać kwadratowa) lub nowsze (formuła pierwiastkowa) elementy rezerwuaru.

4.3. Estymacja predykcyjna

W przypadku gdy niestacjonarność badanego strumienia danych wynika z istnienia uformowanego trendu, warto wprowadzić do modelu elementy prognozowania statystycznego. Dla każdego nowego elementu rezerwuaru x_i, sukcesywnie – od chwili jego uzyskania – budowany jest w tym celu liniowy model predykcyjny; użyta została metoda wygładzania wykładniczego umożliwiająca łatwe uaktualnienie modelu o nowe obserwacje [7]. Do dalszych obliczeń wykorzystywana jest wartość współczynnika kierunkowego powyższego modelu liniowego, dla *i*-tego elementu w chwili *t*, oznaczana przez $a_{2,t,i}$. Jej wielkość (zwłaszcza znak) pozwala na uwzględnienie dynamiki estymatora jądrowego, charakteryzującego zmiany rozkładu strumienia danych w otoczeniu elementu x_i.

Zdefiniujmy funkcję reprezentującą (uciąglone) zmiany estymatora jądrowego. Niech zatem dla ustalonego t dana będzie funkcja $g_t\colon \mathbb{R}^n \to \mathbb{R}$ określona wzorem

$$g_t(x) = \frac{1}{m}\sum_{i=1}^{m} a_{2,t,i} K(x, x_i, h) \ , \tag{17}$$

którą utożsamić można z estymatorem pochodnej czasowej (!) estymatora gęstości badanego strumienia danych. Wprowadźmy następnie współczynnik

$$w_i^{**} = 1 + \beta_i \, sgmKPSS \quad \text{dla } i = 1, 2, \dots, m \ , \tag{18}$$

gdzie $sgmKPSS$ charakteryzuje stopień niestacjonarności, natomiast $\beta_i \in [-1, 1]$ jest parametrem zdefiniowanym dla każdego elementu rezerwuaru. Określmy ich wartość jako

$$\beta_i = \beta_0 \cdot \frac{g_t(x_i)}{\bar{g}_t} \quad \text{dla } i = 1, 2, \dots, m \ , \tag{19}$$

przy czym

$$\bar{g}_1 = 1 \, , \quad \bar{g}_t = \max_{i=1,2,\dots,m_t} |g_t(x_i)| \quad \text{dla } t = 2, 3, \dots \ , \tag{20}$$

natomiast współczynnik $\beta_0 \in [0, 1]$ określa intensywność oddziaływania wprowadzanej niniejszym predykcji. W przypadku niestacjonarności, wartości mniejsze od 1/3 skutkują słabym oddziaływaniem predykcji, a większe niż 2/3 okazują się jako zbyt intensywne. Ostatecznie proponowane jest

$$\beta_0 = \frac{2}{3} \ , \tag{21}$$

a zatem maksymalna wartość z dopuszczalnego przedziału, co intensyfikuje w założonym zakresie wpływ funkcji predykcji.

4.4. Elementy nietypowe

Dzięki wyznaczeniu aktualnej gęstości rozkładu badanego strumienia danych nietrudno jest wyznaczyć obserwacje nietypowe w sensie elementów rzadko występujących. Zastosowano metodę przedstawioną w pracy [8]. Podobnie jak poprzednio, wprowadźmy zatem współczynniki

$$w_i^{***} = 1 + \gamma_i \cdot sgmKPSS \quad \text{dla } i = 1, 2, \dots, m \ , \tag{22}$$

gdzie $sgmKPSS$ charakteryzuje stopień niestacjonarności, natomiast parametr $\gamma_i \in [-1, 1]$ definiuje się w następujący sposób:

$$\gamma_i = \begin{cases} \frac{g_t(x_i)}{\bar{g}_t} & \text{gdy} \quad x_i \text{ jest nietypowy} \\ 0 & \text{gdy} \quad x_i \text{ jest typowy} \end{cases} , \tag{23}$$

przy czym funkcje g_t oraz $\bar{g}_t$ zostały wprowadzone w poprzednim podrozdziale formułami (17), (19) oraz (20). Wybrana metoda identyfikacji elementów rzadkich wymaga sprecyzowania proporcji r elementów nietypowych rezerwuaru w stosunku do jego liczności. Naturalnym jest taki wybór, który wykrywać będzie więcej elementów rzadkich, gdy strumień jest niestacjonarny (co wskazuje na obszary istotnych zmian) i mniej w warunkach niestacjonarności (gdzie takie elementy często są przypadkowe). Przyjęto zatem

$$r = 0.01 + 0.09\, sgmKPSS \ . \tag{24}$$

W praktyce oznacza to, że *r* elementów rezerwuaru zostanie uznanych za nietypowe, a odpowiadające im współczynniki w_i^{***}, będą – dzięki formułom (20) i (23) – większe dla tych, które związane są z trendami wzrostowymi i mniejsze dla elementów nietypowych należących do obszarów recesywnych.

4.5. Końcowe uwagi i sugestie

Ostatecznie, jeżeli używane są wszystkie procedury opisane w podrozdziałach 4.2-4.4, to współczynniki w_i stosowane w ważonej postaci estymatora jądrowego (11) powinny być przyjęte w postaci

$$w_i = w_i^* \cdot w_i^{**} \cdot w_i^{***} \quad \text{dla } i = 1, 2, \dots, m \ . \tag{25}$$

Jeśli któraś z tych procedur – dezaktualizacja, predykcja lub obsługa elementów nietypowych – miałaby

być pominięta, to z powyższego wzoru powinien być usunięty odpowiedni element w_i^*, w_i^{**} lub w_i^{***}. Dla wyrazistości interpretacji, każdy z nich zmienia się w jednakowym zakresie od 0 do 2, a poziom 1 ma charakter neutralny. Wszystkie zmieniają się w sposób ciągły, co upłynnia zmiany w czasie wykresu gęstości rozkładu $\hat{f}$. W trywialnym przypadku, gdy żadna z powyższych procedur nie jest stosowana, należy przyjąć $w_i \equiv 1$. Postać multiplikatywna wzoru (25) wynika z interpretacji rozważanego zagadnienia: duże wartości współczynników w_i powinny charakteryzować te elementy, które spełniają możliwie wszystkie warunki sformułowane w sekcjach 4.2-4.4, a nie jedno lub dwa z nich, co predestynowałoby addytywną postać formuły (25) lub postacie pośrednie.

Wartościowych spostrzeżeń dotyczących czułości modelu predykcyjnego jest sygnał śledzący TS oparty na wartościach błędów prognozowania z jednostkowym wyprzedzeniem, a ściślej definiowany jako iloraz estymatorów wartości oczekiwanej takich błędów oraz ich odchylenia standardowego [6]. Małe wartości bezwzględne sygnału śledzącego $|TS|$, zwłaszcza mniejsze od 0.2, lub okazjonalnie nawet od 0.3, wskazują na poprawność modelu predykcyjnego, w tym doboru czułości modelu predykcji. Inny cenny – poza sygnałem śledzącym TS – sprawdzian poprawności działania procedury DEDSTA można otrzymać obserwując czy liczność rezerwuaru m jest bliska m_0 w okresach względnej stacjonarności, jak również czy w przypadku szybkich zmian, wartość parametru m sukcesywnie maleje do m_{min}, odpowiednio do ich intensywności.

Można również skomentować przypadek wielowymiarowego strumienia danych. Tak zwane przekleństwo wielowymiarowości sprawia, że krytycznym parametrem staje się liczność analizowanego zbioru, a w przypadku algorytmu DEDSTA parametry m_0 oraz m_{min}. Teoretycznie w przypadku wielowymiarowym do zachowania tej samej dokładność estymacji w punkcie zero, w miarę wzrostu wymiaru n, wartości parametrów m_0, m oraz m_{min} powinny być pomnożone przez czynnik 4^{n-1} [10, tabela 4.2]. Spowolniłoby to jednak możliwości adaptacyjne procedury. Warto zatem rozważyć czy zwiększenie liczności jest konieczne dla potrzeb konkretnego zastosowania, czy można pozostawić wartości tych parametrów niezmienione, godząc się z utratą jakości estymacji i w konsekwencji traktując estymator jądrowy bardziej jako wskaźnik struktury rozkładu, a nie ścisłe oszacowanie jego gęstości.

Warto też zwrócić uwagę, że procedura dopuszcza dowolną postać wynikowego jądra wielowymiarowego, produktowe lub radialne, oraz metodę wyznaczania parametru lub parametrów wygładzania, gdyż operuje na wartościach wynikowego estymatora niezależnie od postaci tych metod. Zmniejszenie wartości parametru wygładzania ułatwia obserwację zmian zachodzących w czasie w miarę wzrostu wartości zmiennej niezależnej t.

5. WYNIKI BADAŃ SYMULACYJNYCH

Liczne i wszechstronne badania symulacyjne przeprowadzone zostały z użyciem zarówno syntetycznych, jak i rzeczywistych (meteorologicznych) danych strumieniowych. Pierwsze z nich, dzięki znajomości z góry założonego estymowanego rozkładu, umożliwiły ilościowe wyznaczenie błędów estymacji i w konsekwencji dokonanie szczegółowej analizy polepszania jej jakości, wprowadzanej przez każdy z zaprojektowanych modułów opracowanego algorytmu DEDSTA. W tym celu ponownie analizowane były syntetyczne strumienie danych o zróżnicowanej dynamice niestacjonarności. Ostatecznym narzędziem weryfikacji poprawności estymacji były badania gęstości rozkładu rzeczywistych strumieni danych złożonych z pomiarów meteorologicznych, dla różnych klimatów, a zatem z odmiennymi dynamikami zmian.

Wszystkie omawiane poniżej przebiegi przedstawione są na stronie internetowej dedykowanej tej rozprawie [11], złożone w dwóch opcjach: co 10 i każdy krok. W celu uzyskania płynnego obrazu proponowane jest ściągnięcie plików na dysk komputera – ich przeglądanie bezpośrednio ze strony internetowej może być zbyt wolne i skutkować brakiem płynności ruchu.

Parametr wygładzania wyznaczany był metodą *plug-in*, w przypadku wielowymiarowym odrębnie dla każdej współrzędnej. Otrzymywane tym sposobem wartości były stosunkowo małe, co pozwoliło wyeksponować dynamikę zmian zachodzących w miarę upływu czasu.

Jako reprezentatywny dla uwarunkowań zaprojektowanej metody przyjęto następujący jednowymiarowy strumień danych, ze zmianami dynamiki w postaci naprzemiennych warunków stacjonarnych i niestacjonarnych, o sukcesywnym i gwałtownym charakterze będący sumą składowej deterministycznej oraz losowej, przy czym pierwsza z nich dana jest wzorem

$$X_t = \begin{cases} 0 & \text{dla} \quad t = 1 \\ X_{t-1} + p_{t-1} & \text{dla} \quad t = 2, 3, \dots, 10{,}000 \end{cases}, \tag{26}$$

gdzie

$$p_t = \begin{cases} 0.001 & \text{dla} \quad 1 \le t < 2{,}000 \\ 0.01 & \text{dla} \quad 2{,}000 \le t < 6{,}000 \\ 0 & \text{dla} \quad 6{,}000 \le t < 8{,}000 \\ 1 & \text{dla} \quad t = 8{,}000 \\ 0 & \text{dla} \quad 8{,}000 < t < 10{,}000 \end{cases}, \tag{27}$$

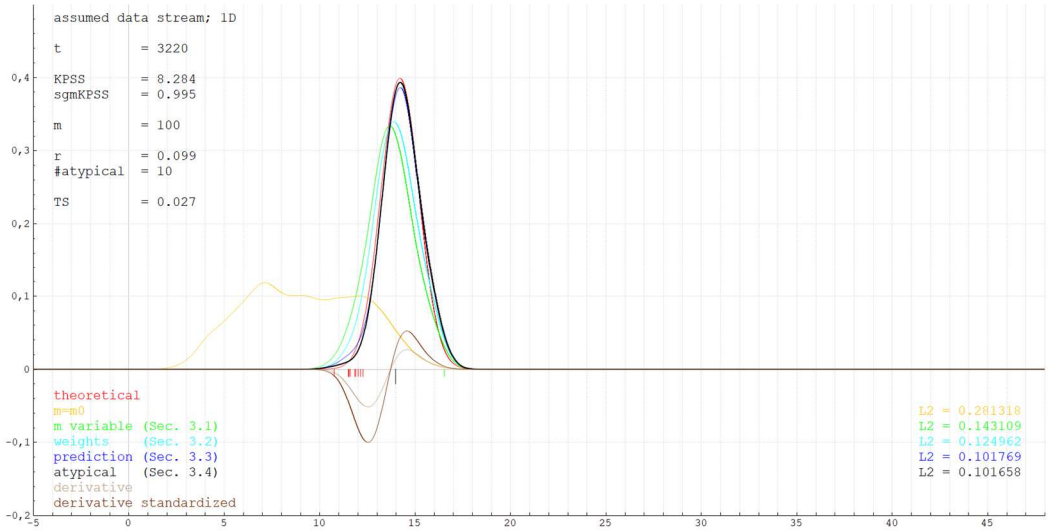
natomiast składowa losowa dla każdego $t = 1,\ 2, \dots, 10{,}000$ ma rozkład normalny standardowy $N(0, 1)$, o nieskorelowanych w czasie wartościach. Warto zwrócić uwagę, że w chwili $t = 8{,}000$ występuje skok jednostkowy. (Na stronie [11] równania (26) i (27) oznaczane są zgodnie z numeracją dysertacji jako (6.1)-(6.2).)

I tak, w czasie początkowego okresu (tj., gdy $1 \leq t < 2{,}000$), o charakterze niestacjonarnym, co stanowi wartą podkreślenia niedogodność, następuje konsolidacja algorytmu. Następnie (w czasie $2{,}000 \leq t < 6{,}000$) ma miejsce istotny wzrost dynamiki strumienia danych – warto zwrócić uwagę, że po 100 krokach wartość oczekiwana rozkładu zmienia się o odchylenie standardowe składowej losowej. W kolejnym okresie ($6{,}000 \leq t < 8{,}000$) gwałtownie pojawiają się warunki stacjonarne, po czym (w chwili $t = 8{,}000$) występuje nagła zmiana w postaci skoku jednostkowego. Na końcu ($8{,}000 < t \leq 10{,}000$) proces staje się stacjonarny. Powyższe elementy, w szczególności takie połączenia zmian dynamiki, stanowią duże wyzwanie dla opracowywanej metody.

W pierwszym okresie konsolidacji ($1 \leq t < 2{,}000$) przede wszystkim ustalane są początkowe wartości parametrów procedury. Algorytm poprawnie wykrywa niedużą niestacjonarność i od momentu zebrania odpowiednio dużej liczby obserwacji, wielkość rezerwuaru m była nieco mniejsza niż przyjęta wartość maksymalna $m_0 = 1{,}000$ (wahała się w zakresie 500-900). W drugim okresie ($2{,}000 \leq t < 6{,}000$), pojawia się istotny trend, który został prawidłowo wykryty, a liczność rezerwuaru uległa istotnej redukcji do $m = m_{min} = 100$. W konsekwencji, także dzięki działaniu funkcji predykcji, estymator nadążał za rozkładem wzorcowym (26)-(27), nierzadko nawet wyprzedzając zmiany wzorca. Po nastaniu w kolejnym okresie ($6{,}000 \leq t < 8{,}000$) warunków stacjonarności, zostały one prawidłowo rozpoznane, a liczność rezerwuaru zwiększyła się do $m = m_0 = 1{,}000$. Następujący w chwili $t = 8{,}000$ skok jednostkowy również został odpowiednio wykryty. W celu szybszej reakcji na jego skutki, liczność rezerwuaru m zmniejszyła się chwilowo do około 200, po czym w kolejnym okresie stabilizacji ($8{,}000 < t \leq 10{,}000$) sukcesywnie wzrastała do $m = m_0 = 1{,}000$.

Na rysunku 1 przedstawiono przykładowe estymatory otrzymane z zastosowaniem procedury DEDSTA dla syntetycznego strumienia danych (26)-(27). Kolorem czerwonym oznaczono gęstość teoretyczną, założoną w generatorze. Żółta krzywa przedstawia estymator zbudowany na pełnym oknie danych, czyli dla stałego $m = m_0 = 1{,}000$. Wynik ten traktowany jest jako referencyjny. Estymator zielony wprowadza zmienną liczność rezerwuaru, zgodnie z tezami podrozdziału 4.1. Estymator z dodatkową dezaktualizacją wag, opisaną w sekcji 4.2, został nakreślony kolorem jasnoniebieskim. Granatową krzywą zaprezentowano estymator wyposażony nadto w moduł predykcji, opisany w sekcji 4.3. Kompletny estymator, uzupełniony o obsługę elementów nietypowych (czyli ze wszystkimi modułami z sekcji 4.1-4.4), został przedstawiony kolorem czarnym. Na rysunku 1 widoczne są ponadto elementy odosobnione, reprezentujące wygasające trendy, które zaznaczono czerwonymi liniami poniżej osi odciętych, a także te stowarzyszone z wstępującymi tendencjami, oznaczone liniami zielonymi. W przypadku opcji każdego kroku (a nie co 10), nieco dłuższa czarna linia poniżej osi odciętych wskazuje najnowszą obserwację. Piaskowa i brązowa krzywa przedstawiają

odpowiednio estymator pochodnej (przemnożonej przez dobrany dla czytelności wykresów współczynnik 10,000) oraz jego standaryzowany odpowiednik, którego ekstremum wartości bezwzględnej znajduje się na ± 0.1, w zależności od znaku. Powyższe kolory poszczególnych estymatorów zostały także opisane w lewym-dolnym rogu wykresu. Odpowiadające kolejnym estymatorom uśrednione wartości indeksów L^2 znajdują w prawym-dolnym rogu obrazu.



Rysunek 1. Estymatory otrzymane w 3.220 kroku za pomocą procedury DEDSTA.

Otrzymane na rysunku 1 wartości wskaźnika L^2 zostały zebrane w tabeli 1. Kolejne kolumny tabeli odpowiadają estymatorowi z dodatkowym – względem poprzedniej – modułem, opisanym w nagłówku kolumny. Widać, że każda wprowadzana zmiana poprawiała wyniki estymacji. Należy jednak zaznaczyć, iż wartość wskaźnika otrzymana przez wprowadzenie modyfikacji wag elementów nietypowych, niewiele różni się od uzyskanej z poprzedniego modułu predykcji. W szczególnych przypadkach mogą być nawet nieco (rzędu części promili) większe. Wynika to z faktu, iż w przypadku modułu elementów nietypowych poprawa jakości estymacji następuje w obszarach mających wyjątkowo mały wpływ na kwadratowy wskaźnik całkowy L^2. Należy jednak zaznaczyć, że już sama wizualizacja na wykresie wykrytych elementów nietypowych z zaznaczeniem czy reprezentują one tendencje wzrostowe czy regresywne, potencjalnie może stać się cenna dla wszechstronnej analizy eksperckiej, uwzględniającej czy nierzadko wręcz nakierowanej na dynamikę zmian. Ostatecznie, wobec podstawowej, referencyjnej metody ruchomego okna o stałej liczności, uzyskano poprawę wskaźnika L^2 rzędu ponad 60%. Dla dodatkowego zobrazowania powyższych wyników, w tabeli 2 zawarto procentowe zmiany wartości wskaźników jakości względem uzyskanych z użyciem poprzednich modułów.

Tabela 1. Porównanie jakości estymatorów dla syntetycznego strumienia danych.

	stałe m (referencyjny)	zmienne m (sekcja 4.1)	dezaktualizacja (sekcja 4.2)	predykcja (sekcja 4.3)	elementy nietypowe (sekcja 4.4)
wskaźnik L^2	0.2813	0.1431	0.1250	0.1018	**0.1017**
zmiana procentowa	100 %	49.13 %	55.56 %	63.81 %	**63.85 %**

Tabela 2. Poprawa jakości estymatorów względem poprzednich modułów.

	stałe m (referencyjny)	zmienne m (sekcja 4.1)	dezaktualizacja (sekcja 4.2)	predykcja (sekcja 4.3)	elementy nietypowe (sekcja 4.4)
zmiana procentowa	---	50.87 %	12.85 %	18.56 %	0.001 %

Procedura DEDSTA została także porównana z innymi metodami, reprezentującymi różne ujęcia estymacji gęstości rozkładu danych strumieniowych, oparte na odmiennych koncepcjach: falkową [5], tzw. jądra klastrowe [4], a także sieci samoorganizujące [1]. Wyniki porównań dla szeregu (26)-(27) zawarto w tabeli 3. Procedura DEDSTA znacząco przewyższała powyższe metody, aczkolwiek należy zauważyć, że badania prowadzone były właśnie z warunkami, dla których została ona skonstruowana.

Tabela 3. Zestawienie najlepszych wyników uzyskanych z użyciem odmiennych metod estymacji.

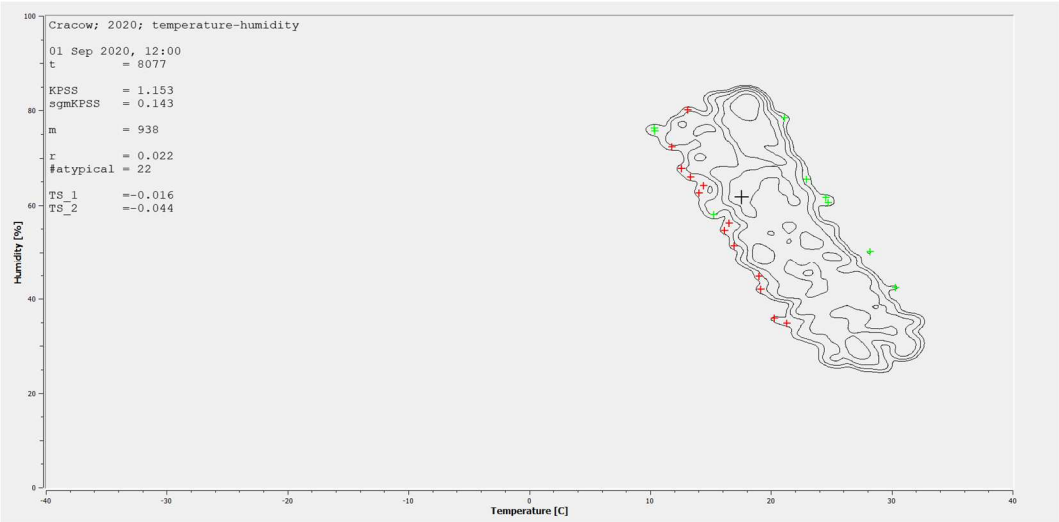
	metoda falkowa	jądra klastrowe	sieci samoorganizujące	DEDSTA
wskaźnik L^2	0.149	0.300	0.446	**0.073**

(W reakcji na uwagi zawarte w recenzjach dokonano także porównania z najnowszym znalezionym w literaturze algorytmem ALoKDE z 2021 roku [3]. W tym przypadku uzyskano wartość wskaźnika L^2 wynoszącą 0.141, a więc także znacząco gorszą niż otrzymaną z użyciem procedury DEDSTA.)

Ostatnim elementem przedstawianych tu wyników weryfikacji numerycznej były strumienie danych, złożone z pomiarów meteorologicznych, dotyczących temperatury i wilgotności względnej w

Minneapolis, Rio de Janeiro oraz Krakowie, a zatem trzech miast o całkowicie odmiennym klimacie. Ograniczenie wymiarowości do dwóch wynika z możliwości odpowiedzialnej graficznej ilustracji wyników. W przypadku 2-wymiarowym wyniki przedstawiono w postaci map konturowych z logarytmicznym układem wartości poziomic; najniższym zaznaczonym poziomem jest 0.0003125, a każdy kolejny był dwukrotnie większy od poprzedniego. Podobnie jak poprzednio wyniki zostały załączone na stronie internetowej [11] w dwóch opcjach: co 10 i każdy krok. W celu uzyskania płynnego obrazu proponowane jest ściągnięcie plików na dysk komputera. Otrzymane estymatory można interpretować jedynie na podstawie analizy jakościowej, a zwłaszcza potencjalnej przydatności do szeroko pojętych zagadnień analizy i eksploracji danych, przede wszystkim zgodności z doświadczeniem w zakresie corocznych zmian pogodowych, gdyż w tym przypadku gęstość teoretyczna oczywiście nie jest znana. Warto zwrócić uwagę nie tylko na same mapy konturowe, ale także na kierunki pojawiania się elementów nietypowych związanych z tendencjami wzrostowymi (plusy zielone) i malejącymi (plusy czerwone), a w przypadku jednowymiarowym także na wartości pochodnych czasowych (krzywe o kolorze beżowym), wskazującymi kierunki zmian z pewnym wyprzedzeniem. W przypadku opcji „w każdym kroku", czarny krzyż oznacza obserwację aktualną.

Otrzymane wyniki są łatwo interpretowalne, w pełni zgodne z doświadczeniem. Zmiany były płynne, a kolor elementów nietypowych prawidłowo opisywał dynamikę ruchu. Na rysunku 2 pokazano przykładowy 2-wymiarowy wyniki dla Krakowa na początku meteorologicznej jesieni. Widać znaczne rozciągnięcie rozkładu wzdłuż kierunku mała temperatura – duża wilgotność (dni deszczowe) oraz duża temperatura – mała wilgotność (dni pogodne). Elementy nietypowe wyraźnie wskazują na wzrostową tendencję temperatury (koniec lata) oraz powiększanie wilgotności (początek jesieni).



Rysunek 2. Estymator otrzymany za pomocą procedury DEDSTA dla łącznego rozkładu temperatury i wilgotności względnej w Krakowie na początku meteorologicznej jesieni, 1 września 2020 roku.

6. PODSUMOWANIE

Celem niniejszej dysertacji było opracowanie i przedstawienie nowatorskiej nieparametrycznej metody estymacji gęstości rozkładu danych strumieniowych. Ostateczna procedura, nazwana DEDSTA, może być użyta zarówno w warunkach niestacjonarnych jak i stacjonarnych, przy czym mogą one następować naprzemiennie, a w pierwszym przypadku intensywność zmian może się istotnie różnicować. Koncepcja tej procedury została oparta na nieparametrycznej metodzie estymatorów jądrowych, co pozwala wyznaczyć gęstość praktycznie dowolnego występującego w praktyce rozkładu, bez konieczności formułowania o nim dodatkowych, często arbitralnych założeń. Ponadto stosowano elementy prognozowania statystycznego, umożliwiające efektywne nadążanie za zachodzącymi zmianami, a także identyfikowane były elementy nietypowe ze wskazaniem, które z nich związane są z nowopowstałymi tendencjami, a które ze ustępującymi.

Struktura procedury DEDSTA składa się z niezależnych modułów: wyznaczania liczności rezerwuaru, dezaktualizacji elementów, predykcji, a także obsługi elementów nietypowych (w sensie rzadkiego ich występowania). Najpierw wyznaczana jest liczba z przedziału [0, 1] charakteryzująca stopień niestacjonarności strumienia danych, która stanowi potem podstawę przy wyznaczaniu parametrów niezbędnych do działania powyższych modułów. Każdy z nich może być potencjalnie pominięty albo ewentualnie dostosowany do specyficznych uwarunkowań rozważanego zagadnienia, aczkolwiek należy podkreślić, że przedłożona metoda w swej podstawowej postaci jest kompletna i nie wymaga dodatkowych badań koncepcyjnych. Wobec wszystkich parametrów zostały zaproponowane wartości standardowe, po czym przestawiono analizę wrażliwości procedury na zmiany ich wartości. Każda z użytych składowych proponowanej tu metody ma liniową lub kwadratową złożoność obliczeniową względem bieżącej liczności rezerwuaru, nieprzekraczającej 1000. Zarówno czas obliczeń jak i wymagania pamięci nie wykraczają poza możliwości współczesnych systemów komputerowych. Obliczenia każdego kroku trwają około sekundy na standardowym sprzęcie.

Przedstawiona tematyka ma charakter rozwojowy. Tematem dalszych badań będzie zastąpienie koncepcji ruchomego okna, polegającego na wykluczaniu najstarszych elementów, poprzez usuwanie losowe z prawdopodobieństwem zależnym od aktualnego stopnia niestacjonarności. Pozwoli to na uniknięcie całkowitej eliminacji elementów starszych niż bieżąca liczność okna. Innym aspektem przyszłych badań będzie uogólnienie przedstawionej koncepcji o możliwość uwzględnienia atrybutów kategorycznych, a także redukcję wymiarowości współrzędnych ciągłych i kategorycznych, uzależnionej od aktualnych uwarunkowań badanego strumienia danych. Kolejnym przedmiotem prac będzie ujęcie warunkowe, w ramach którego możliwe jest uwzględnienie pomiarów tych wielkości, których aktualna wartość pozwala na istotne uściślenie stosowanego modelu probabilistycznego.

BIBLIOGRAFIA (WYBRANE POZYCJE)

1. Cao Y., He H., Man H.: SOMKE: Kernel Density Estimation Over Data Streams by Sequences of Self-Organizing Maps, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, ss. 1254-1268, 2012.
2. Chacon J.E., Duong T.: "Multivariate Kernel Smoothing and Its Applications", Chapman and Hall/CRC, 2018.
3. Chen Z., Fang Z., Sheng V., Zhao J., Fan W., Edwards A., Zhang K.: Adaptive Robust Local Online Density Estimation for Streaming Data, *International Journal of Machine Learning and Cybernetics*, vol. 12, ss. 1803-1824.
4. Heinz C., Seeger B.: Cluster Kernels: Resource-Aware Kernel Density Estimators Over Streaming Data, *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, ss. 880-893, 2008.
5. Heinz C., Seeger B.: Wavelet density estimators over data streams, *Proceedings of the 2005 ACM symposium on Applied computing*, ss. 578-579, 2005.
6. Holden K., Peel D.A.: On Testing for Unbiasedness and Efficiency of Forecast, *The Manchester School*, vol. 58, ss. 120-127, 1990.
7. Hyndman R.J., Koehler A., Ord J.K., Snyder R.D.: *Forecasting with Exponential Smoothing: The State Space Approach*, Springer, 2009.
8. Kulczycki P., Kruszewski D.: Identification of atypical elements by transforming task to supervised form with fuzzy and intuitionistic fuzzy evaluations, *Applied Soft Computing*, vol. 60, ss. 623-633, 2017.
9. Kwiatkowski D., Phillips P.C.B., Schmidt P., Shin Y.: Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root, *Journal of Econometrics*, vol. 54, ss. 159-178, 1992.
10. Silverman B.W.: *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1986.
11. *Strona internetowa pracy*, https://www.ibspan.waw.pl/~trybotyc/Thesis_Page/main.html, dostęp 9 kwietnia 2022.

25 IV 2023, Tomasz Rybotycki

